Organización de Datos 75.06. Segundo Cuatrimestre de 2018. Examen parcial, segunda oportunidad:

Importante: Antes de empezar complete nombre y padrón en el recuadro. Lea bien todo el enunciado antes de empezar. Para aprobar se requiere un mínimo de 60 puntos (60 puntos = 4) con al menos 20 puntos entre los ejercicios 1 y 2. Este enunciado debe ser entregado junto con el parcial si quiere una copia del mismo puede bajarla del grupo de la materia. En el ejercicio 3 elija 2 de los 4 ejercicios y resuelva única y exclusivamente 2 ejercicios. Si tiene dudas o consulta levante la mano, está prohibido hablar desde el lugar, fumar o cualquier actividad que pueda molestar a los demás. El criterio de corrección de este examen está disponible en forma pública en el grupo de la materia.

"I ignored my destiny once. I cannot do that again. Even for you. I'm sorry, Gamora." - Thanos, Avengers Infinity War

#	1	2	3.1[ ]	3.2[ ]	4	5	6	7	Entrega Hojas:
Corrección									Total:
Puntos	/15	/15	/10	/10	/15	/10	/15	/10	/100

Nombre:
Padrón:
Corregido por:

1) Spotify cuenta con un log de todas las canciones que fueron escuchadas en su plataforma, esta información se encuentra en un RDD que está paralelizado por día, es decir cada día es una partición. Los campos del RDD son los siguientes: (date, user_id, song_id, song_title, artist).

Se cuenta por otro lado con un RDD que asigna "tags" a las canciones, por ejemplo "rock, punk, actual, top-10, acoustic, etc). Una canción puede tener asociados "n" tags. El RDD tiene el formato (song_id, tag).

Cada día se corre un proceso para asignar el tag "rising" a las canciones que se escucharon mas veces el día de hoy que el día de ayer. Estos nuevos tags pasan a formar parte del RDD de tags para futuros usos. Programar en **PySpark usando el API de RDD** lo solicitado.

(15 pts) (***)

2) El dataframe (sales) lista las ventas de productos con los siguientes campos: Dia, Mes, Año, ProductID, Importe(USD). Para un mismo dia, mes y año puede venderse n veces el mismo producto. Por otro lado tenemos una descripción de los productos en el dataframe (products): ProductId, Title, Category, Description. Category puede ser "Men", "Women", "Kids"

Proponer un programa en **Pandas** que permita:

a) Indicar los títulos de los productos de la categoría "Men" para los cuales el Importe de venta supera el promedio mensual de ventas de los productos de la misma categoría. (por ejemplo si el promedio de Abril de "Men" es 120 dolares y un producto se vendió en Abril a 135 dolares lo tenemos que listar). Usar Transform. (*** 7pts)

b) Indicar el top-10 de productos que se vendieron mayor cantidad de días de forma consecutiva. (**** 8 pts)

3) Resolver 2 (dos) y solo 2 de los siguientes ejercicios a elección (si resuelve mas de 2 el ejercicio vale 0 puntos, sin excepciones). En cada caso indicar V o F justificando adecuadamente sus respuestas. **Si no justifica vale 0 puntos sin excepciones.**

a) Ningún algoritmo puede comprimir más del 1% de todos los archivos posibles en al menos 1 byte. (**)(10 pts)

b) Existe al menos un compresor capaz de comprimir el 0.5% de todos los archivos posibles a la mitad de su tamaño. (**) (10 pts)

c) Para cualquier n, existe al menos un string x con longitud n que es incompresible (*) (10 pts)

d) El nivel de compresión luego de concatenar dos strings X e Y es el mismo sin importar el orden porque K(XY) = K(YX) (*) (10 pts)

4) (***) Sea la siguiente matriz de 1's y 0's que indican con un 1 si el documento Si contiene el elemento *i*:

Elemento	S_1	S_2	S_3	S_4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

Se pide:

a. Encontrar, para cada documento, los 3 valores de Minhash utilizando las funciones $h1(x) = 2x + 1 \bmod 6$, $h2(x) = 3x + 2 \bmod 6$, $h3(x) = 5x + 2 \bmod 6$. Explique el paso a paso.(5pts)

b. Con los valores anteriores, estimar la semejanza de Jaccard S(S1,S2) y S(S2,S4) y explicar si esta estimación coincide con la semejanza de Jaccard verdadera.(5pts)

c. ¿Cuáles pares de documentos son candidatos a ser similares (usar b=1 r=3)? (5pts)

5) (10ptos) (***) Dados los siguientes documentos:

Abanico Azul Aro

Arco Aro Avion

Arco Arpa Azul Avioneta

Avion Avioneta

Sabiendo que se construyó sobre ellos un índice invertido utilizando front coding parcial (n=3) y códigos gamma para codificar los punteros, indicar cuantos accesos son necesarios para resolver la consulta "Arco Azul" detallando paso a paso la forma en que se resuelve la misma.

6) (***) Sea la siguiente DVS de una matriz A de rango 2 que contiene calificaciones de películas:

$$A = \begin{bmatrix} .14 & 0 \\ .42 & 0 \\ .56 & 0 \\ .70 & 0 \\ 0 & .60 \\ 0 & .75 \\ 0 & .30 \end{bmatrix} \begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix} \begin{bmatrix} .58 & .58 & .58 & 0 & 0 \\ 0 & 0 & 0 & .71 & .71 \end{bmatrix}$$

U Σ V^T

Asumiendo que las columnas de A son películas ([Matrix, Alien, Star Wars, Casablanca, Titanic]) y las filas son los usuarios ([Joe, Jim, John, Jack, Jill, Jenny, Jane]), se pide:

1. Interpretar qué información brindan los dos autovectores. ¿Puede encontrar dos grupos de usuarios que difieran entre sí por algún motivo? (6pts)
2. La primer columna de U contiene un .14 y un .70. ¿Qué conclusión puede sacar respecto a las calificaciones que realizan Joe y Jack? (6pts)
3. ¿Cree que si hubiéramos utilizado más autovalores y autovectores podríamos aproximarnos aún mejor a la matriz A? Justifique. (3pts)

7) Crear una única visualización que permita representar las tres matrices del ejercicio 6, y que transmita el significado de las mismas. (***) (10 pts)